

IMG/VR: A database of cultured and uncultured DNA Viruses and Retroviruses

Summary

Viruses represent the most abundant life forms on the planet. Recent experimental and computational improvements have led to a dramatic increase in the number of viral genome sequences identified primarily from metagenomic samples. As a result of the expanding catalog of metagenomic viral sequences, there exists a need for a comprehensive computational platform integrating all these sequences with associated metadata and analytical tools. Here we present IMG/VR (<https://img.jgi.doe.gov/vr/>), the largest publicly available database of isolate reference DNA viruses with over 260,000 computationally identified viral contigs from more than 6,000 ecologically diverse metagenomic samples. Approximately half of the viral contigs are grouped into genetically distinct *quasi*-species clusters, enabling prediction of microbial host(s) for 20,000 viral sequences, revealing 14 microbial phyla previously unreported to be infected by viruses. Viral sequences can be queried using a variety of associated metadata, including habitat type and geographic location of the samples, or taxonomic classification according to hallmark viral genes. IMG/VR has a user-friendly interface that allows users to interrogate all integrated data and interact by comparing with external sequences, thus serving as an essential resource in the viral genomics community.

The resource

The Integrated Microbial Genomes with Virus (IMG/VR) database is an integrated viral analysis system within the IMG with Microbiome samples (IMG/M) data management system.

IMG/VR provides the largest integration of viral sequences with associated metadata and allows users to explore these data to decipher biogeographical and habitat distribution patterns of viral species as well as traveling across all the identified hosts putatively infected with viral sequences. In addition, users can compare and analyze their sequences against IMG/VR's data (including viral protein family models, viral cluster and singleton information, distribution patterns of similar viral sequences across the globe, percent of known and unknown genes per sequence, and information regarding viral taxonomy and putative viral-host(s)), integrated with a variety of analytical tools.

We anticipate that IMG/VR will become a reference resource for sequence analysis of viral genomes and viral contigs derived from metagenomic samples.

How to use IMG/VR

Browsing iVGs and mVCs via Viral Datasets

The search functionality in IMG/VR is similar to that in the IMG/M system. All isolate viral genomes (iVGs) can be accessed via “*Quick Genome Search*” (by typing the virus name or taxon identifier “Taxon OID”) or “*Find Genomes*” tab (selecting *viruses* in “*Genome Browser*” or “*Genome Search*” tools) (**Fig. 1**).

The predicted mVCs are stored as metagenome *scaffolds* and they remain under their corresponding metagenome datasets (i.e. metagenome “Taxon OID”). Thus, metagenome “Taxon OIDs” can also be accessed the same way that any iVG and specific mVCs can be retrieved from the “*Scaffold Search*” tool of the “*Find Genomes*” tab (**Fig. 1**).

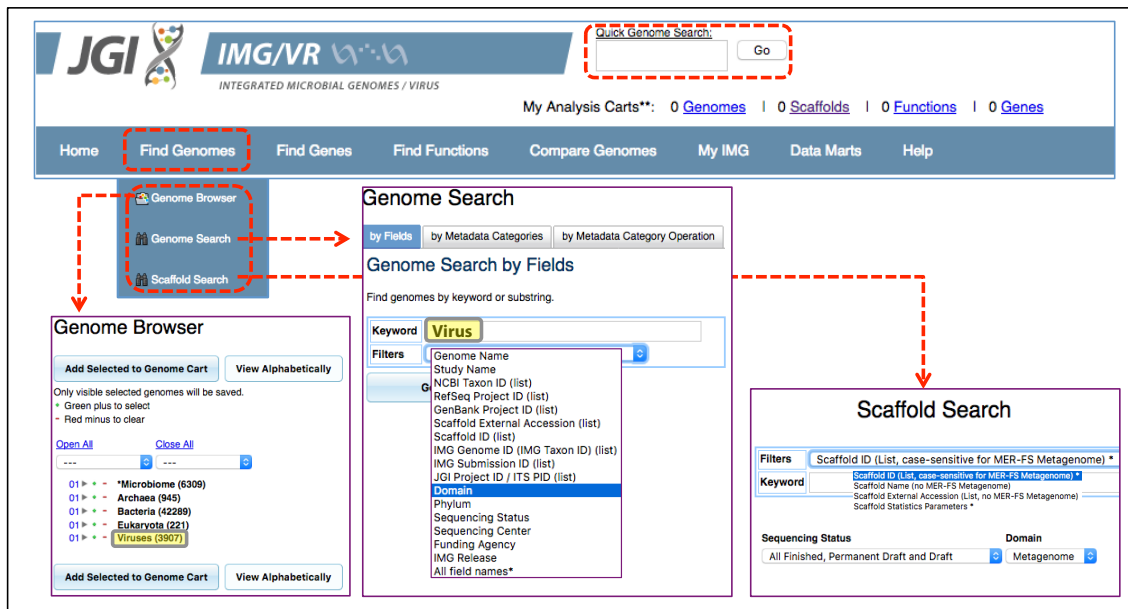


Figure 1. General IMG/VR search functionality. Basic search tools from IMG/M’s top menu bar (dashed red boxes) can be used to access the viral content of IMG/VR. “*Quick Genome Search*” at the top menu can be used to query specific viral names, taxon identifiers or keywords. Alternatively, all isolate viral content can be retrieved from the “*Find Genome*” tab, either selecting *Viruses* (boxed in grey with yellow background) from the “*Genome browser*” display (bottom left panel) or searching for *Virus* (boxed in grey with yellow background) in the *Domain* filter of the “*Genome Search*” tool (bottom central panel). To search for metagenomic viral contigs users need to access first the metagenomic sample (using any of the above tools). Then the “*Scaffold Search*” tool can be used to select specific scaffolds (bottom right panel).

In order to further facilitate the identification and selection of viral sequences in IMG/VR, all iVGs and mVCs can be accessed from the left panel table (IMG Viral Content) available from the entry page (*Home* tab) (**Fig. 2a**). This entry point enables browsing all viral

datasets in the context of their associated samples and corresponding metadata, e.g. habitat type or depth of the metagenome sample from which a viral sequence was identified (**Fig. 2b**). This table provides information about the total number of viral contigs per sample in IMG, allowing a quick identification of the samples with the largest number of viruses. Similar to other tables in IMG, the results can be exported in a tab-delimited text format compatible with other tools for metagenomics analysis, as well as R and Microsoft Excel (**Fig. 2b**). By clicking on the “*Viral Contig Count*” number from the previous table, users can examine the list of viral contigs from individual samples (**Fig. 2c**). The information displayed for a selected contig or group of contigs includes: scaffold identifier (*Scaffold ID*), gene count per contig (*Gene Count*), contig length (*Sequence Length bp*), guanine and cytosine content (*GC Content*), percent of genes per contig covered with viral protein families (*Perc VPFs*), viral species name identifier (*Viral Cluster*; detailed in “**Sequence grouping**” section and **Supplementary data**), predicted host and method of prediction (*Host Detection*; detailed in “**Host-virus identification**” section), taxonomic assignment at different levels based on POGs (**Supplementary data**), and the putative retrovirus sequences (**Supplementary data**).

The screenshot displays the IMG/VR (Integrated Microbial Genomes / Virus) web interface. At the top, there is a navigation bar with links like Home, Find Genomes, Find Genes, Find Functions, Compare Genomes, My IMG, Data Marts, and Help. Below this, a section titled 'IMG Viral Content' shows counts for Isolate Viruses (3907), Metagenomic Viral Contigs (mVCs) (264413), and Total Viral Datasets (268320). A dashed red oval highlights the 'Total Viral Datasets' link. To the right, a table titled 'All Viral Datasets' lists various studies with columns for Study name, Taxon OID, Genome Name, Habitat Type, and Viral Contig Count. A red dashed line connects the 'Total Viral Datasets' link to the 'Viral Contig Count' column. Below the table, a section titled 'Metagenome Viral Contigs' shows a list of contigs with columns for Scaffold ID, Gene Count, Sequence Length, GC content, perc VPFs, and Viral Cluster. A red dashed line connects the 'Viral Contig Count' column to the 'Viral Cluster' column. The interface also includes a 'Quick Genome Search' bar, a 'My Analysis Carts' section, and a 'Download VPF Models' button.

Figure 2. Browsing IMG/VR viral datasets. (a) Total counts and access to the list of viral sequences from isolate viruses, metagenomic viral contigs, or their combination (dashed red oval). (b) Detailed table from “*Total Viral Datasets*” link displaying study and sample name, Taxon OID, habitat information, and number of metagenomic viral contigs. (c) List of viral metagenomic contigs found in a single sample. Columns in (b) and (c) can be sorted by clicking on the column header, and different filters can be used for specific searches (black boxes). Tables can be also exported in a tabular format by using the *Export* button (grey box with yellow background).

A

Engineered	14091
Bioreactor	640
Bioremediation	1548
Biotransformation	605
Food production	11
Lab enrichment	1146
Lab synthesis	71
Solid waste	536
Unclassified	280
Wastewater	9254

B

Environmental	203190
Air	98
Aquatic	193080
Terrestrial	10012

C

Host-associated	42271
Annelida	134
Arthropoda	1712
Birds	385
Fish	46
Fungi	2
Human	30839
Mammals	2334
Microbial	30
Mollusca	74
Plants	6578
Porifera	20
Tunicates	14
Unclassified	3

Environmental Terrestrial														
Add Selected to Genome Cart				Select All	Clear All									
Iter column - Viral Contig Count Filter text Apply														
Export	Page 1 of 5		1	2	3	4	5	100/2, 100/2, 100/2						
Select	Domain	Status	Study Name	Taxon ID	Genome Name	Ecosystem Ecosyst / Category / Type / Subtype / Specific	Depth	Habitat Type	Habitat From GOLD	Viral Contig Count				
<input type="checkbox"/>	D	D	Soil microbial communities from Rifle, Colorado - Rifle Deep Mine (RDM) - 12 Rifle CSP/Plank high/012, ASSEMBLY_DATE: 20140222	330000040	Environmental	Terrestrial	Soil	Loam	Unclassified	Terrestrial (Soil)	Soil	343		
<input type="checkbox"/>	D	D	Agricultural soil microbial communities from Utah and Georgia to study Nitrogen management - Poultry litter 2014	330000079	Environmental	Terrestrial	Soil	Unclassified	Unclassified	Terrestrial (Soil)	Agricultural soil	255		
<input type="checkbox"/>	D	D	Deep subsurface shale carbon reservoir microbial communities from Ohio and West Virginia, USA	330000063	Environmental	Terrestrial	Deep subsurface	Unclassified	Unclassified	Om	Terrestrial (other)	Deep subsurface	233	
<input type="checkbox"/>	F	F	Soil microbial communities from Great Plains (Iowa, Continuous Corn soil (Iowa, Continuous Corn soil, Feb 2012 Assem MSU haeey-qand)	330000068	Environmental	Terrestrial	Soil	Unclassified	Unclassified	Grasslands	4 inch with litter removed	Terrestrial (Soil)	Soil	195
<input type="checkbox"/>	D	D	Corn, switchgrass and miscanthus rhizosphere microbial communities from Kellogg Biological Station, Michigan, USA	330000044	Environmental	Terrestrial	Soil	Loam	Unclassified	Agricultural Soil	Terrestrial (Soil)	Corn, switchgrass and miscanthus rhizosphere	182	

Figure 3. Accessing metagenomic viral contigs via associated environmental metadata. (a) Distribution of metagenomic viral contigs per *ecosystem* and *ecosystem category* information of samples according to GOLD classification. When a category is selected (e.g. *Terrestrial* samples -boxed in dashed red) a new table is displayed. **(b)** Detailed information of the selected *Terrestrial* samples. The total number of metagenomic viral contigs per sample (boxed in dashed red) can be viewed. Columns can be sorted by clicking on the column header, and different filters can be used for specific searches (black box). The table can be also exported in a tab-delimited text format by using the *Export* button (greyen box with yellow background). **(c)** Number of mVCs per *Habitat Type* category of the sample where the mVCs were found.

Browsing mVCs via geographic location or human body site metadata

Viral contigs can be viewed based on the geographic coordinates of a corresponding sample. This functionality is available primarily for environmental metagenomes and allows the selection of samples with specific location via “*Marker Clusterer for Google Maps*”, a javascript API utility library that creates and manages per-zoom-level clusters for large amounts of markers in Google Maps. Ultimately, as a user zoom in the map, The a list of viral contigs that belong to those a sample(s) can be retrieved by clicking on a map pin and selecting the count next to the metagenome of interest for that location (**Fig. 4a**).

Additionally, all viral contigs identified in samples from the human body can be displayed by clicking on the “*Show Human Body Sites*” button (**Fig. 4b**). This option allows access to viral contigs derived from samples of any of the 5 main human body sites (nose, mouth, skin, intestine, and vagina), together with general statistics of these viruses per body site (**Fig. 4c**). From the default *Human Body Sites* summary table users can select all mVCs from a particular sample site or only those with a putative host (**Fig. 4d**).

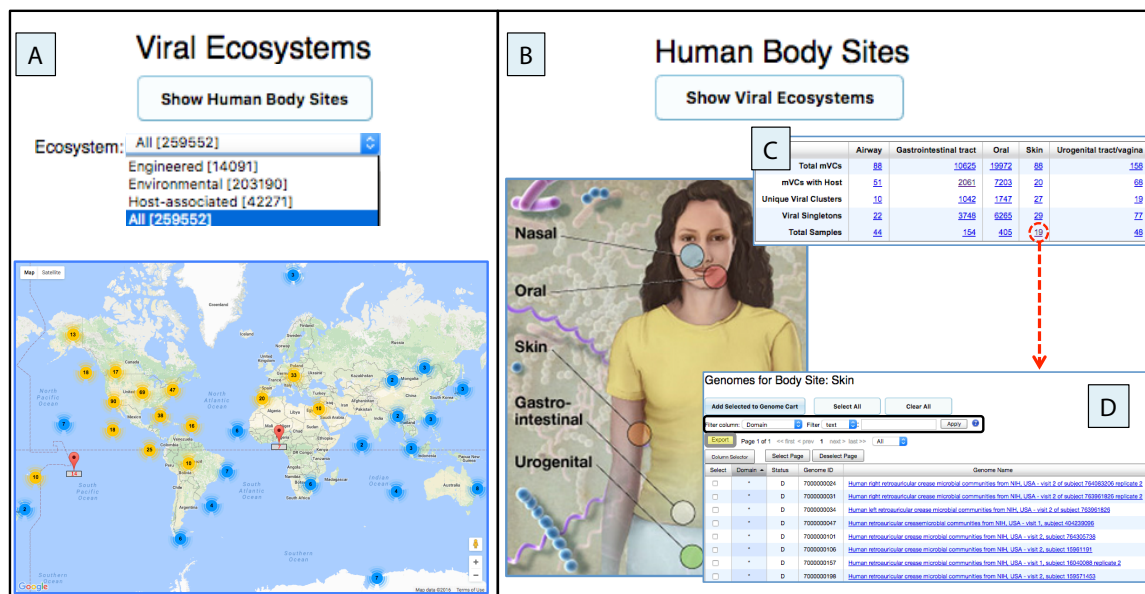


Figure 4. Maps of samples containing viral contigs from environmental and human-associated metagenomes. (a) World interactive Google Map with a geographic location of metagenomic samples. All samples can be selected together or only those from any of the three major ecosystems. Map pins (in red) represent location counts of viral contigs and may contain multiple samples. Map pins are grouped into clusters and clusters themselves into larger clusters (bold number with a coloured halo based on number of members within the cluster) according to the Google Map javascript API utility library. As you zoom into any of the cluster locations, the number on the cluster decreases, and you begin to see the individual markers on the map. Zooming out of the map consolidates the markers into clusters again. (b) Human Body image showing the five body sites with available samples. All the metagenomic viral contigs identified in each body site can be accessed from the circles in the image. (c) Table provides information about mVC clusters/singletons, number of samples, and viral contigs with a host. (d) List of human skin samples with viral contigs. Columns can be sorted by clicking on the column header. Different filters can be used for specific searches (black box). This table can be exported by using the *Export* button (grey box with yellow background). Human body image credit: NIH Medical Arts and Printing.

Browsing viral clusters and viral singletons

Viral clusters and singletons together represent the entire viral diversity within IMG/VR. A total of 39,701 viral clusters and 122,665 singletons are available from the left panel on IMG/VR's entry page (**Fig. 5a**). Together, these represent 162,366 viral *quasi*-species identified numerically with the prefix “vc_” or “sg_” depending if they belong to a viral cluster or remain as a singleton.

By clicking on the viral cluster or singleton identifiers the users can obtain information about the number of members in the cluster (“*Viral Contig Count*”), the number of samples in which they were found (“*Sample Count*”), the number of independent projects these samples belong to (“*Study Count*”), the proposed host (when detected, “*Host*”), and the sample's habitat (“*Habitat Type*”) (**Fig. 5b**).

By clicking on a single viral cluster, all members of the cluster are displayed with several related metadata, including the number of genes per viral contig, contig length, GC content, host assignment, and taxonomic information (**Fig. 5c**).

Finally, the total number of metagenomic viral sequences that can be assigned to a host (at the lowest possible taxonomic level) by projecting the host-virus information onto a viral cluster, is also presented. There are 13,947 in this category, whereby in the majority of the cases the virus-host link is at genus or species level. The microbial genera infected with the highest number of viral contigs are *Streptococcus*, *Veillonella*, *Fusobacterium*, and *Prevotella* s (**Fig. 6d**). In ~9% of all assignments, the host connection is at a higher taxonomy range (ranging from family to phylum).

All the information from all the tables can be independently accessed by clicking on their corresponding links or could be exported in an excel tab-delimited text format by using the “*Export*” button.

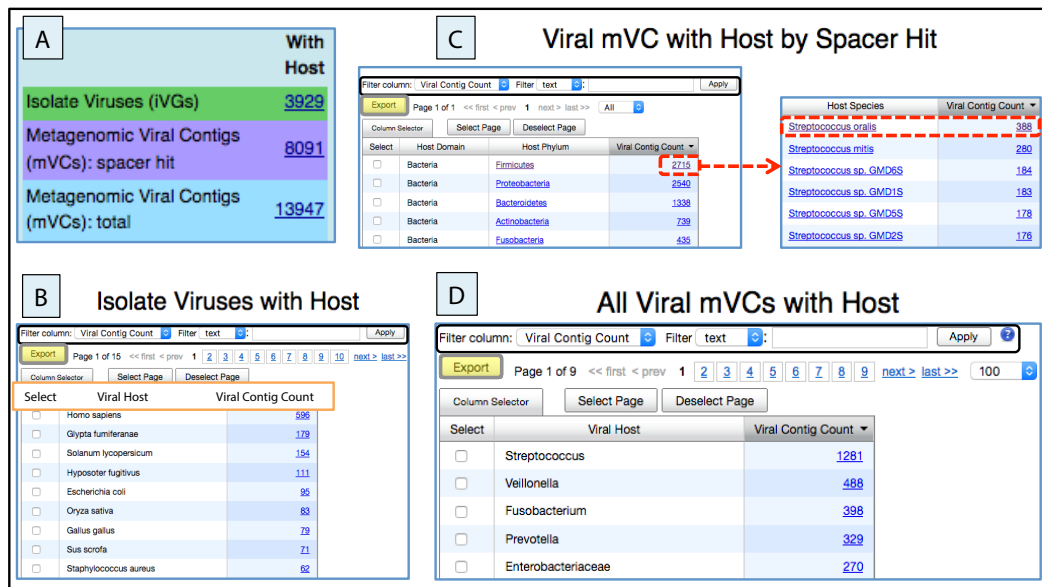


Figure 6. Viral data sets with host assignment. (a) Number of isolate viruses or metagenomic viral contigs with a predicted host. (b) “*Isolate viruses with host*” table sorted by the hosts infected by the highest number of viral genomes. (c) Top microbial host species containing metagenomic viral contigs. (d) Microbial hosts (at different taxonomic levels) with the highest number of metagenomic viral contigs assigned. Columns in (b), (c), and (d) can be sorted by clicking on the column header and different filters can be used for specific searches (black boxes). Tables can be also exported in a tab-delimited text format by using the *Export* button (grey box with yellow background).

Sequence search

Users can compare their sequences against the sequence data integrated into IMG/VR. Specifically, the sequences of all the viral contigs and all the spacer sequences from the isolate genomes can be queried by using the “*Viral/Spacer Blast*” option at the bottom of the home page (**Fig. 7a**). Both queries can be selected from “*Blast Database*” and rely on nucleotide BLAST searches (23) with customizable e-value identity cutoffs (**Fig. 7b**).

Matches against the viral database generate a list of viral sequences with a significant alignment based on the selected thresholds. These *subject* sequences can be directly accessed or selected-to-be-added to the *Scaffold Cart*, where their associated metadata are also provided. Similarly, matches of external viral sequences against the spacer database generate a list of host(s) which contain a CRISPR-spacer sequence with a significant alignment based on the selected cutoffs. These putative host(s) can be further explored by clicking on the host identifier. This redirects the user to detailed information of the spacer: source taxon name, location of the spacer within the CRISPR array, and spacer sequence (**Fig. 7c**).

Figure 7. Viral searches against IMG/VR databases. (a) Location of *blast* tool in IMG/VR (dashed red box). (b) User interface to *blast* sequences. Exclusively nucleotide sequences can be queried currently in the system. Sequence(s) must be added into the blank area. Users can *blast* their sequence(s) against any of the 2 databases integrated into IMG/VR: “*Viral Sequence*” or “*Viral Spacer*” and customize the e-value cutoff. (c) Example of a *blast* output of an external partial viral sequence (Streptococcus phage 858) against the spacer database. When a sequence hit (purple box) is selected, a new panel (“*Viral Spacer*”) is displayed showing details of the sequence spacer and the putative corresponding microbial host.